Algorithm 1 OPNP - Optimal Parameter and Neuron Pruning

Require: A trained model $f(\boldsymbol{x};\boldsymbol{\theta})$, training samples $\{\boldsymbol{x}_i\}_{i=1}^m$, pruning percentage ρ^w_{max}, ρ^w_{min}, ρ^o_{min}, ρ^o_{max}, test sample $\{\boldsymbol{x}_k\}_{k=1}^n$;
Ensure: Energy score of test sample $\{E(\boldsymbol{x}_k)\}_{k=1}^n$

for $i = 1$ to m **do**
  compute gradient $g(\boldsymbol{x}_i)$;
end for
$\mathbf{M} = \frac{1}{m}\sum_{k=1}^{m}|g(\boldsymbol{x}_k)|$;
$\mathbf{O}_i = \frac{1}{K}\sum_{p=1}^{K}\mathbf{M}_{ip}, \quad i = 1, \cdots, L$;
Get threshold Ω^w_{min}, Ω^w_{max} by ranking $\mathbf{M}$;
Get threshold Ω^o_{min}, Ω^o_{max} by ranking $\mathbf{O}$;
$\mathbf{W}[\mathbf{M} > \Omega^w_{max}] = 0$ and $\mathbf{W}[\mathbf{M} < \Omega^w_{min}] = 0$;
for $k = 1$ to n **do**
  Compute the pre-logit embedding $h(\boldsymbol{x}_k)$;
  $h(\boldsymbol{x}_k)[\mathbf{O} > \Omega^o_{max}] = 0$ and $h(\boldsymbol{x}_k)[\mathbf{O} < \Omega^o_{min}] = 0$;
  $f(\boldsymbol{x}_k) = \mathbf{W} \cdot h(\boldsymbol{x}_k) + \boldsymbol{b}$;
  $E(\boldsymbol{x}_k) = -\log\sum_{i=1}^{K}\exp(f_i(\boldsymbol{x}_k))$;
end for
Return $\{E(\boldsymbol{x}_k)\}_{k=1}^n$

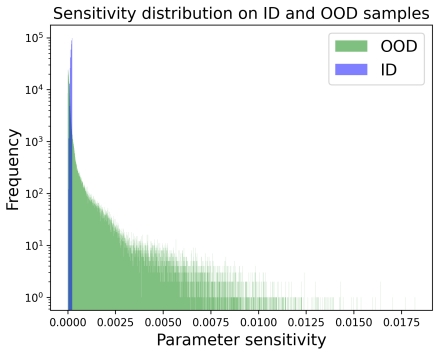

Figure 1: Sensitivity distribution on ID and OOD samples. The top 20% sensitive parameters (measure on ID set) are illustrated. The average sensitivity on OOD samples is 0.00024 and the average sensitivity on ID samples is 0.00018.

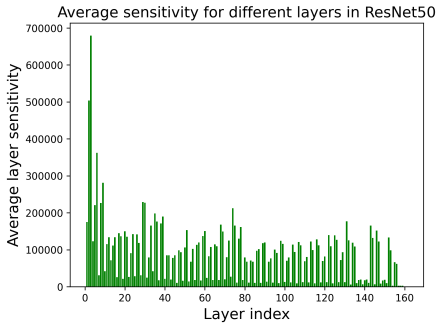

Figure 2: Average parameter sensitivity for different layers in ResNet50.